Combinatorial protein design

Jeffery G Saven

Combinatorial protein libraries permit the examination of a wide range of sequences. Such methods are being used for *de novo* design and to investigate the determinants of protein folding. The exponentially large number of possible sequences, however, necessitates restrictions on the diversity of sequences in a combinatorial library. Recently, progress has been made in developing theoretical tools to bias and characterize the ensemble of sequences that fold into a given structure – tools that can be applied to the design and interpretation of combinatorial experiments.

Addresses
Department of Chemistry, University of Pennsylvania, 231 South 34 Street, Philadelphia, Pennsylvania 19104, USA;
e-mail: saven@sas.upenn.edu

Current Opinion in Structural Biology 2002, **12**:453–458

0959-440X/02/$ – see front matter

Introduction

The discovery and design of novel proteins can lead to new, potentially practical proteins and can also enhance our understanding of protein biochemistry. Designing well-structured, soluble proteins is difficult, however, because of their complexity. Such proteins are large (tens to hundreds of amino acid residues) and have many variables that specify the folded state, including sequence, backbone topology and sidechain conformation. Design involves identifying those sequences that fold into a given structure from a huge ensemble of possible sequences. This search is aided, in part, by the large degree of consistency seen in folded proteins. On average, a folded structure is well packed, hydrophobic residues are sequestered from solvent and most potential hydrogen bond interactions are satisfied. This consistency, however, is often complex, may have little simplifying symmetry and involves predominantly noncovalent interactions. Such interactions are some of the most difficult to accurately quantify. As such, estimating the free energies associated with mutation or structural ordering remains a subtle area of computational research. Nonetheless, many molecular potentials do contain a 'best parameterization' of many of the interatomic interactions and forces that we know are important for stabilizing proteins. In some cases, such potentials have been used with striking success in protein design [1••]. Given that these potentials are necessarily approximate, however, one promising approach is to use the partial information contained in these functions in a probabilistic manner. A probabilistic or statistical approach is also appropriate for characterizing the full variability of sequences that fold to a common structure, because there are likely to be an enormous number of such sequences. Such statistical methods can be applied in 'shotgun' approaches to *de novo* protein design. Combinatorial experiments create and assay many sequences in order to overcome shortcomings in our understanding of folding or other molecular properties. Even though combinatorial methods can address large numbers of sequences (10^4–10^{12}), these numbers are still infinitesimal in comparison to the numbers of possible sequences (e.g. $20^{100} \approx 10^{130}$ for a 100-residue protein). Thus, methods for winnowing and focusing sequence space are a vital component of combinatorial protein design. Herein, I briefly discuss combinatorial methods for full sequence design. I also review recent theoretical developments in characterizing sequence ensembles — developments that can be applied to the design and interpretation of combinatorial experiments.

Directed protein design

There has been much effort — and success — in developing computational methods for 'directed' protein design. By 'directed protein design', I mean the identification of a sequence (or a small set of sequences) that is likely to fold into a predetermined backbone structure. Each such sequence can then be synthesized to confirm its folded structure and other molecular properties. Early efforts in design identified proteins with substantial order, but not necessarily well-defined tertiary structure [2]. Because an enormous number of sequences are possible even for small proteins (<50 residues), computational methods have dramatically accelerated successful design. Typically, such methods are implemented as an optimization process, whereby amino acid identity and sidechain conformation are varied in order to optimize a scoring function that quantifies sequence/structure compatibility. Exhaustive searching of all m^N possible sequences (where m is the number of different amino acid types or 'states' per residue and N is the number of residues in a target protein structure) is feasible only if a small number of residues N are allowed to vary or if the number of amino acids m is greatly reduced. If, in the optimization process, the different sidechain conformations (rotamer states) of each amino acid are also considered (see [3]), the complexity of the search increases still further, because m, the number of possible 'states' per residue, increases by a factor of ten or more. Although complete enumeration is typically not feasible, sequence space can be sampled in a directed manner in order to find optimal (or nearly optimal) sequences. Stochastic methods, such as genetic algorithms or simulated annealing, involve searching sequence space in a partially random fashion; on average, the search progressively moves toward better scoring (lower energy) sequences [4,5]. The partially random nature of the search permits escape from local minima in the sequence/rotamer landscape. Using a simplified model, the Takada and Tamura groups have included information about unfolded structures (negative design) in a stochastic search for a sequence with a 'funneled conformational energy landscape' [6]. One

47-residue three-helix bundle protein so selected has CD and NMR spectral features of folded proteins (W Jin, O Kambara, H Sasakawa, A Tamura, S Takada, personal communication). When applied to atomically detailed representations, the stochastic methods focus primarily on repacking the interior of a structure with hydrophobic residues [7] and have been applied to the wild-type structures of 434 Cro [8], ubiquitin [9], the B1 domain of protein G [10•], the WW domain [1••] and helical bundles [11,12]. Although, in many cases, these methods have identified experimentally viable sequences [1••,13], stochastic search methods need not identify global optima [14•]. For potentials comprising only site and pair interactions, elimination methods such as 'dead end elimination' can find the global optimum [14•,15–17]. Such methods successively remove individual amino acid rotamer states that cannot be part of the global optimum until no further states can be eliminated. The Mayo group applied such methods to automate the full sequence design of both a 28-residue zinc finger mimic [18] and, after predetermining hydrophobic and polar sites, a 51-residue homeodomain motif [19•]. The group has also redesigned portions of a variety of proteins [20–22]. Functional properties such as metal binding or catalysis may also be included as elements of the design process [23,24•]. The elements and algorithms of directed protein design have been the subject of several recent reviews [1••,25,26•].

Despite some striking successes, computational methods for directed design have limitations with respect to both identifying folding sequences and characterizing the features of protein sequences that share a common structure. Stochastic methods, such as simulated annealing or genetic algorithms, can be applied to large proteins and permit many sites to be varied simultaneously, but the computational times and resources required for such calculations are extensive, even for small proteins. When used as optimization methods, directed approaches will necessarily be sensitive to the energy or scoring function used. All energy functions in use in protein design, however, are necessarily approximate and uncertainties in the energy function may not merit the search for global optima. Furthermore, many naturally occurring proteins are not optimized. In fact, most proteins are only marginally stable (e.g. ΔG^{o} <10 kcal/mol for folding) [27]. In addition, sequences that function, for example, those that bind another molecule, need not be the global optimum with respect to structural stability. Although stochastic methods can sample such suboptimal sequences, in general an exponentially large number of them will be possible and such sampling will be time consuming. Thus, it is important to develop methods complementary to those used for directed protein design — methods that reveal the features of sequences that are likely to fold into a particular structure but that may not be structurally 'optimal'. Such computational methods will have application to a new class of protein design studies, combinatorial experiments, in which large numbers of proteins may be simultaneously synthesized and screened.

Combinatorial design

Combinatorial design provides a complementary approach to directed design for understanding sequence/ structure compatibility and discovering novel sequences that fold into a specific structure. Combinatorial methods are powerful tools for cases in which we have an incomplete understanding of molecular properties. In protein combinatorial design experiments, large numbers of sequences (libraries) are screened for evidence of folding into a predetermined structure. A combinatorial experiment has two key elements: creating a library with a desired degree of diversity and assaying for sequences with 'protein-like' properties in terms of their structure or function. Depending upon how the diversity is generated and assayed, experiments of this type can explore a large number of sequences, up to 10^{12} [28•]. Certainly, such methods can be used to discover 'hits', that is, a few sequences that are especially stable or that are unusually strong in their function or binding properties. In addition, combinatorial experiments readily generate a sequence ensemble. Thus, using combinatorial experiments, we can potentially 'expand the protein sequence database' and the diversity of these additional sequences will be at the control of the researcher. Features important to folding (and other properties) may be explored in a way that is decoupled from the evolutionary requirements of nature's proteins. For example, these methods have been used to identify helical proteins [29–31], ubiquitin variants [32], self-assembled protein monolayers [33], proteins with amyloid-like properties [33], metal-binding peptides [34] and stable interhelical oligomers [35]. Several excellent reviews of combinatorial experiments have appeared recently [36,37,38•,39••].

The complexity of combinatorial experiments implies that limitations must be placed on the sequences, because the number that can be created and screened (10^{6}–10^{12}) is infinitesimal compared to the number possible (e.g. 10^{130}). Limitations on sequence properties are often guided by qualitative chemical considerations, but quantitative computational methods will be helpful in designing and interpreting combinatorial experiments.

The Hecht group has probed the extent to which the patterning of hydrophobic and hydrophilic residues can successfully reduce complexity in combinatorial design. While maintaining the periodicity of α helices and β sheets in particular tertiary structures, such patterning is applied in order to expose hydrophilic residues to solvent and to sequester hydrophobic residues in the interior of the protein. Early targets were helical proteins; a fiducial 74-residue four-helix bundle was the template structure [40]. Such a structure has more than $20^{74} \approx 10^{96}$ possible sequences. After binary patterning, five hydrophobic and six hydrophilic amino acids were permitted at 24 interior and 36 exterior positions, respectively, thus reducing the total number of possible sequences to 10^{41}. From a protein library consistent with this binary patterning, a set of 50 correctly expressed sequences was selected for further

study. Around half of the 50 sequences isolated are protein-like in many respects [30], including their thermal denaturation [41]. About half the isolated sequences also bind heme [29] and many of these display carbon monoxide binding [42•] or peroxidase activity [43]. This is surprising given that such functions were not part of the design or selection of the sequences. In a second-generation design, the group added six residues to each of the four helices of one of the most protein-like sequences. The additional residues were combinatorially patterned, as in the original experiment [39••]. For these 102-residue sequences, the free energies of folding are increased 2–3-fold and the NMR data suggest well-determined structures. Using binary patterning of hydrophobicity consistent with an amphilic β sheet [44], the Hecht group has also identified proteins that aggregate to form amyloid fibrils [45] and crafted monomeric β proteins by introducing a nonpolar lysine mutation at the 'edge' strand of the target β sheet [46••].

Despite the striking results from hydrophobic patterning, more detailed methods for library design are merited. Many of the hydrophobically patterned sequences that appear well structured are not sufficiently soluble for NMR structure determination [46••] and, as a result, little is known concerning their structures at the atomic scale. Not all of the α-helical sequences exhibit the sharp thermal transition seen in natural proteins (usually associated with a large ΔH of folding). Such sequences may not possess well-packed interiors [41]. In natural proteins, the side-chains of most interior residues are well determined, as opposed to the variability that is obtained using hydrophobic patterning alone and that is observed in many *de novo* designed proteins [13,18]. A more fine-grained dictation of the amino acid identities is probably necessary for obtaining libraries that are rich in sequences with well-defined structures. Moreover, a more detailed specification of amino acid identities yields fewer sequences than hydrophobic patterning alone and further reduces the complexity of the library.

Theories of combinatorial libraries

Surveying the complete sequence landscape of proteins seems, at first glance, intractable to both experiment and computation. In addition to the enormous number of possible sequences, many examples exist in nature of dissimilar sequences folding to essentially the same structure. Hence, sequence properties are nontrivial and proteins sharing a common structure can be nonlocal in sequence space. Nonetheless, computational methods permit us to estimate the properties, particularly the amino acid probabilities, of sequences consistent with a target structure.

Repeated use of directed search methods can estimate the properties of an ensemble of sequences. Desjarlais and co-workers have used independent runs of their sequence prediction algorithm across an ensemble of closely related structures all consistent with a particular fold (JR Desjarlais *et al.*, personal communication). For each structure, an optimal 'nucleating' sequence is identified and subsequently the sequence/rotamer variability is explored throughout the structure. The method identifies effective reduced partition sums for each sequence/rotamer state and amino acid probabilities may be obtained at each residue position. The number of sequences decreases with stability, so the degree of complexity can be tuned by varying a cutoff in the effective free energies of the sequences. The method has been used to identify sequences consistent with the fold of a WW domain, a small β-sheet protein [1••], some of which are currently being experimentally characterized.

The amino acid frequencies can also be determined directly, using a statistical theory of combinatorial libraries [47,48••,49••]. Ideas from statistical mechanics are used to address the number and composition of sequences that are consistent with a particular backbone structure. The theory addresses the whole space of available compositions, not just the small fraction that is accessible to experiment and to computational enumeration and sampling. The theory takes as input a target backbone structure and a scoring or energy function for quantifying sequence/structure compatibility. Global and local features can be prespecified using constraints on the sequences. For example, such constraints can be used to determine the energy the sequences assume in the target structure, the patterning of amino acids and the number of each amino acid present (composition). The theory yields estimates of both the number of sequences consistent with these constraints and the amino acid probabilities at each residue position. These residue-specific probabilities are the most probable such set and are determined — as in statistical mechanics — by maximizing an effective entropy, whereby this maximization is subject to constraints. Just as in thermodynamics, the judicious use of constraints can be used to reduce the entropy or the number of possible sequences. Thus, these methods provide a systematic means to focus the library, winnowing numbers such as 10^{130} to numbers that are experimentally manageable, for example, 10^6. The theory agrees well with exact results obtained with lattice models of proteins [47,48••]. This method has been extended to realistic representations of proteins, in which the effects of sidechain packing are included in an atom-based manner [49••]. The calculated sequence probabilities of the immunoglobulin light chain binding domain of protein L are in agreement with the frequencies observed in combinatorial phage display experiments [50,51]. These statistical methods have several advantages. They may be applied to much larger proteins (N>100 residues) and permit much larger sequence variation than many directed methods. They are sufficiently rapid that many backbone structures may be considered and those features that are robust with respect to minor structure modifications may be identified. Importantly, such methods provide perhaps the most natural input for a combinatorial experiment, the probabilities of the amino acids at each position among the sequences of a library. These amino acid

probabilities can also be used to identify specific amino acid sequences, which can then be synthesized; a consensus sequence comprising the most probable amino acid at each site can be selected or the probabilities can be used to bias a stochastic search for viable sequences (J Zou, JG Saven, unpublished data).

If the energy of the target state is one of the constraints, the statistical method reduces to an effective mean field theory. Mean field theories have seen extensive application in physical science and in biomolecular theory [52], and to protein evolution and natural sequence variability ([53]; H Kono, JG Saven, unpublished data). Voigt *et al.* [14•] have compared mean field theories with directed search methods for identifying ground state sequence/rotamer combinations in protein design. They found that, although often more rapid, mean field theories do not always identify such ground states. Interestingly, Voigt *et al.* applied the mean field theory to large proteins (subtilisin E and T4 lysozyme) to determine local site entropies, s_i, where $\exp(s_i)$ quantifies the effective number of amino acids allowed at residue i in a structure [54••,55]. Sites with large values of s_i, those most tolerant to mutation [56], are likely to support substitutions that improve stability or function when *in vitro* evolution experiments are used to explore sequence space [37]. For such experiments, the mutation rate is low enough that multiple mutations of strongly interacting sites are rare. Thus, mutations that improve 'fitness' are most likely to accumulate at sites that are the most 'decoupled' from other sites. Such mutations can potentially be targeted for variation in an *in vitro* evolution experiment.

Conclusions

Much recent progress has been seen in the design and discovery of new proteins, and combinatorial approaches are accelerating the pace. Such methods are most useful when our quantitative understanding of important protein properties, such as stability and catalytic activity, is limited. Not only can combinatorial methods be used for discovery but also, more deeply, they can inform our understanding of protein properties by generating and assaying whole ensembles of sequences. Traditionally, advances in structural biology have come from examining the structures of naturally occurring proteins, but, with combinatorial experiments, an enormous diversity of sequences can be generated at the control of the researcher. Detailed questions can be addressed, such as the utility of hydrophobic patterning or of predetermining particular sites for amino acid variation. Theory and simulation will continue to aid the design and interpretation of combinatorial experiments. Such methods will also facilitate the exploration of what is possible with the amino acids: how diverse is the set of all possible sequences that fold to a particular structure and what structures not yet seen in nature can be crafted with the amino acids? Such methods will perhaps have an even more profound impact on designing nonbiological foldamers [57••], structures about which we have much less empirical information than we do about biopolymers.

Acknowledgements

The author acknowledges support from the National Science Foundation (CHE 98-16497 and CHE 99-84752). JGS is a Cottrell Scholar of Research Corporation and is an Arnold and Mabel Beckman Foundation Young Investigator.

References and recommended reading

Papers of particular interest, published within the annual period of review, have been highlighted as:

- • of special interest
- •• of outstanding interest

1. Kraemer-Pecore CM, Wollacott AM, Desjarlais JR: **Computational protein design.** *Curr Opin Chem Biol* 2001, 5:690-695.

•• This is a compact but excellent review on recent progress in computational methods for protein design. The authors also discuss recent efforts in designing the WW domain, a small β protein.

2. Bryson JW, Betz SF, Lu HS, Suich DJ, Zhou HX, O'Neil KT, DeGrado WF: **Protein design: a hierarchic approach.** *Science* 1995, 270:935-941.

3. Dunbrack R: **Rotamer libraries.** *Curr Opin Struct Biol* 2002, 12:in press.

4. Shakhnovich EI, Gutin AM: **A new approach to the design of stable proteins.** *Protein Eng* 1993, 6:793-800.

5. Jones DT: ***De novo* protein design using pairwise potentials and a genetic algorithm.** *Protein Sci* 1994, 3:567-574.

6. Onuchic JN, Luthey-Schulten Z, Wolynes PG: **Theory of protein folding: the energy landscape perspective.** *Annu Rev Phys Chem* 1997, 48:545-600.

7. Hellinga HW, Richards FM: **Optimal sequence selection in proteins of known structure by simulated evolution.** *Proc Natl Acad Sci USA* 1994, 91:5803-5807.

8. Desjarlais JR, Handel TM: ***De-novo* design of the hydrophobic cores of proteins.** *Protein Sci* 1995, 4:2006-2018.

9. Johnson EC, Lazar GA, Desjarlais JR, Handel TM: **Solution structure and dynamics of a designed hydrophobic core variant of ubiquitin.** *Structure* 1999, 7:967-976.

10. Jiang X, Farid H, Pistor E, Farid RS: **A new approach to the design**
• **of uniquely folded thermally stable proteins.** *Protein Sci* 2000, 9:403-416.

The authors use a novel scoring function for the design of hydrophobic interiors. In addition to steric interactions, the function includes parameterizations of changes in the heat capacity and the conformational entropy upon folding. Simulated annealing was used to optimize the score. The backbone and exterior residue identities were constrained. In tests on two small proteins, in which 10–11 interior residues were varied, the native sequence was regenerated, as well as the sequences of known stable variants. Interestingly, previously designed sequences with low stability and weak cooperativity were not identified. In larger proteins tested, in which 32 and 63 residues were varied, sequence/rotamer combinations close to native were identified.

11. Jiang X, Bishop EJ, Farid RS: **A *de novo* designed protein with properties that characterize natural hyperthermophilic proteins.** *J Am Chem Soc* 1997, 119:838-839.

12. Bryson JW, Desjarlais JR, Handel TM, DeGrado WF: **From coiled coils to small globular proteins: design of a native-like three-helix bundle.** *Protein Sci* 1998, 7:1404-1414.

13. Walsh STR, Cheng H, Bryson JW, Roder H, DeGrado WF: **Solution structure and dynamics of a *de novo* designed three-helix bundle protein.** *Proc Natl Acad Sci USA* 1999, 96:5486-5491.

14. Voigt CA, Gordon DB, Mayo SL: **Trading accuracy for speed:**
• **a quantitative comparison of search algorithms in protein sequence design.** *J Mol Biol* 2000, 299:789-803.

The authors carefully compared different methods for sequence design, including simulated annealing, genetic algorithms, mean field methods and dead end elimination (DEE). DEE most reliably finds global minima, but the authors also note that the method may be limited to 30 amino acid sites for which full amino acid variability is permitted. The authors extrapolate the results to regimes to which DEE cannot be applied. They find that both mean field and annealing approaches perform best with core residues and less reliably with residues that are fully or partially solvent exposed.

15. Gordon DB, Mayo SL: **Radical performance enhancements for combinatorial optimization algorithms based on the dead-end elimination theorem.** *J Comput Chem* 1998, 19:1505-1514.

16. Gordon DB, Mayo SL: **Branch-and-terminate: a combinatorial optimization algorithm for protein design.** *Structure* 1999, 7:1089-1098.

17. Pierce NA, Spriet JA, Desmet J, Mayo SL: **Conformational splitting: a more powerful criterion for dead-end elimination.** *J Comput Chem* 2000, 21:999-1009.

18. Dahiyat BI, Mayo SL: ***De novo* protein design: fully automated sequence selection.** *Science* 1997, 278:82-87.

19. • Marshall SA, Mayo SL: **Achieving stability and conformational specificity in designed proteins via binary patterning.** *J Mol Biol* 2001, 305:619-631.
The authors combined binary patterning with atomistic sidechain interactions to identify folding sequences of a homeodomain. Preliminary calculations were done using 'generic amino acids' to identify those sites that are most appropriate for polar or hydrophobic residues. Subject to this binary patterning, a directed search was then performed using dead end elimination. Interestingly, the authors identified an optimal binary patterning, whereby adding or subtracting hydrophobic residues adversely affects folding to stable monomers.

20. Malakauskas SM, Mayo SL: **Design, structure, and stability of a hyperthermophilic protein variant.** *Nat Struct Biol* 1998, 5:470-475.

21. Strop P, Mayo SL: **Rubredoxin variant folds without iron.** *J Am Chem Soc* 1999, 121:2341-2345.

22. Shimaoka M, Shifman JM, Jing H, Takagi L, Mayo SL, Springer TA: **Computational design of an integrin I domain stabilized in the open high affinity conformation.** *Nat Struct Biol* 2000, 7:674-678.

23. DeGrado WF, Summa CM, Pavone V, Nastri F, Lombardi A: ***De novo* design and structural characterization of proteins and metalloproteins.** *Annu Rev Biochem* 1999, 68:779-819.

24. • Bolon DN, Mayo SL: **Enzyme-like proteins by computational design.** *Proc Natl Acad Sci USA* 2001, 98:14274-14279.
The authors designed non-native enzymatic activity into a thioredoxin fold. The authors computationally identified promising active sites on the scaffold. The sequence was designed to stabilize the transition state of a hydrolysis reaction. The enzymes so designed had activity well above background.

25. Street AG, Mayo SL: **Computational protein design.** *Structure* 1999, 7:R105-R109.

26. • Saven JG: **Designing protein energy landscapes.** *Chem Rev* 2001, 101:3113-3130.
The author reviews recent progress in protein design from the perspective of the energy landscape theory of folding. In the context of theory, models and real systems, different issues involved in design are discussed, including target structures, energy functions, foldability criteria, search methods and the size of the amino acid alphabet.

27. Gromiha MM, Uedaira H, An J, Selvaraj S, Prabakaran P, Sarai A: **ProTherm, thermodynamic database for proteins and mutants: developments in version 3.0.** *Nucleic Acids Res* 2002, 30:301-302.

28. • Keefe AD, Szostak JW: **Functional proteins from a random-sequence library.** *Nature* 2001, 410:715-718.
A fascinating study on a random search for 'function' among random amino acid sequences. Using combinatorial methods the authors have pioneered, ATP-binding proteins were selected from a library of 10^{12} sequences.

29. Rojas NRL, Kamtekar S, Simons CT, Mclean JE, Vogel KM, Spiro TG, Farid RS, Hecht MH: ***De novo* heme proteins from designed combinatorial libraries.** *Protein Sci* 1997, 6:2512-2524.

30. Roy S, Ratnaswamy G, Boice JA, Fairman R, McLendon G, Hecht MH: **A protein designed by binary patterning of polar and nonpolar amino acids displays native-like properties.** *J Am Chem Soc* 1997, 119:5302-5306.

31. Roy S, Helmer KJ, Hecht MH: **Detecting native-like properties in combinatorial libraries of *de novo* proteins.** *Fold Des* 1997, 2:89-92.

32. Finucane MD, Tuna M, Lees JH, Woolfson DN: **Core-directed protein design. I. An experimental method for selecting stable proteins from combinatorial libraries.** *Biochemistry* 1999, 38:11604-11612.

33. Xu GF, Wang WX, Groves JT, Hecht MH: **Self-assembled monolayers from a designed combinatorial library of *de novo* beta-sheet proteins.** *Proc Natl Acad Sci USA* 2001, 98:3652-3657.

34. Case MA, McLendon GL: **A virtual library approach to investigate protein folding and internal packing.** *J Am Chem Soc* 2000, 122:8089-8090.

35. Arndt KM, Pelletier JN, Muller KM, Alber T, Michnick SW, Pluckthun A: **A heterodimeric coiled-coil peptide pair selected *in vivo* from a designed library-versus-library ensemble.** *J Mol Biol* 2000, 295:627-639.

36. Zhao HM, Arnold FH: **Combinatorial protein design: strategies for screening protein libraries.** *Curr Opin Struct Biol* 1997, 7:480-485.

37. Giver L, Arnold FH: **Combinatorial protein design by *in vitro* recombination.** *Curr Opin Chem Biol* 1998, 2:335-338.

38. • Hoess RH: **Protein design and phage display.** *Chem Rev* 2001, 101:3205-3218.
A comprehensive review of a commonly used method to generate and display combinatorial libraries of proteins and peptides.

39. •• Moffet DA, Hecht MH: ***De novo* proteins from combinatorial libraries.** *Chem Rev* 2001, 101:3191-3203.
A review of recent work on the *de novo* combinatorial design of proteins, focusing primarily on the pioneering work of the Hecht group.

40. Kamtekar S, Schiffer JM, Xiong HY, Babik JM, Hecht MH: **Protein design by binary patterning of polar and nonpolar amino-acids.** *Science* 1993, 262:1680-1685.

41. Roy S, Hecht MH: **Cooperative thermal denaturation of proteins designed by binary patterning of polar and nonpolar amino acids.** *Biochemistry* 2000, 39:4603-4607.

42. • Moffet DA, Case MA, House JC, Vogel K, Williams RD, Spiro TG, McLendon GL, Hecht MH: **Carbon monoxide binding by *de novo* heme proteins derived from designed combinatorial libraries.** *J Am Chem Soc* 2001, 123:2109-2115.
Heme-assisted binding of a diatomic ligand turns out to be easier to find than expected within a library of sequences patterned to form a four-helix bundle. Eight combinatorially selected heme-binding sequences bind carbon monoxide with an affinity similar to that of myoglobin. The binding properties of the proteins aren't nearly as diverse as those seen among natural heme proteins, but these *de novo* sequences serve as a useful 'reference'.

43. Moffet DA, Certain LK, Smith AJ, Kessel AJ, Beckwith KA, Hecht MH: **Peroxidase activity in heme proteins derived from a designed combinatorial library.** *J Am Chem Soc* 2000, 122:7612-7613.

44. West MW, Beasley JR, Hecht MH: **Collections of *de novo* beta-sheet proteins designed by binary patterning of polar and nonpolar amino acids.** *Protein Eng* 1997, 10:38-38.

45. West MW, Wang WX, Patterson J, Mancias JD, Beasley JR, Hecht MH: ***De novo* amyloid proteins from designed combinatorial libraries.** *Proc Natl Acad Sci USA* 1999, 96:11211-11216.

46. •• Wang WX, Hecht MH: **Rationally designed mutations convert *de novo* amyloid-like fibrils into monomeric beta-sheet proteins.** *Proc Natl Acad Sci USA* 2002, 99:2760-2765.
Previously, the authors had used hydrophobic patterning consistent with β sheets that intermolecularly align 'edge on' and found these sequences did indeed form the amyloid fibrils that were expected. In this paper, they break up these edge-on interactions with a hydrophilic residue (lysine) at each edge of the β sheet. These sequences are indeed monomeric and appear to be well structured according to CD and NMR peak dispersion. These are the first examples of combinatorial β-protein design.

47. Saven JG, Wolynes PG: **Statistical mechanics of the combinatorial synthesis and analysis of folding macromolecules.** *J Phys Chem B* 1997, 101:8375-8389.

48. •• Zou JM, Saven JG: **Statistical theory of combinatorial libraries of folding proteins: energetic discrimination of a target structure.** *J Mol Biol* 2000, 296:281-294.
The authors extend their statistical theory of sequence libraries to include negative design. The sequence space is resolved in multiple dimensions and the number of sequences is characterized according to the folded state energy and stability gap (the difference in energy between the folded state and an ensemble of unfolded conformations). Excellent agreement is observed between theoretical and exact lattice model results for both the numbers of sequences and the monomer probabilities.

49. •• Kono H, Saven JG: **Statistical theory for protein combinatorial libraries. Packing interactions, backbone flexibility, and the sequence variability of a main-chain structure.** *J Mol Biol* 2001, 306:607-628.
A statistical theory of combinatorial libraries developed for combinatorial experiments. The authors used an atom-based potential and rotamer states to identify the sequence probabilities consistent with a particular structure. An effective one-body energy was introduced that relates the hydrophobicity

or solvent-exposure propensity to local β-carbon density. The calculations give good results with regard to sidechain modeling. Calculations were done that are consistent with recent combinatorial experiments on protein L. Generally, the calculations are in good agreement with the observed amino acid frequencies, despite the sampling issues that are always a concern with these comparisons. (Only 20–40 sequences were sequenced in the experiments.)

50. Kim DE, Gu HD, Baker D: **The sequences of small proteins are not extensively optimized for rapid folding by natural selection.** *Proc Natl Acad Sci USA* 1998, **95**:4982-4986.

51. Gu H, Doshi N, Kim DE, Simons KT, Santiago JV, Nauli S, Baker D: **Robustness of protein folding kinetics to surface hydrophobic substitutions.** *Protein Sci* 1999, **8**:2734-2741.

52. Koehl P, Delarue M: **Mean-field minimization methods for biological macromolecules.** *Curr Opin Struct Biol* 1996, **6**:222-226.

53. Dokholyan NV, Shakhnovich EI: **Understanding hierarchical protein evolution from first principles.** *J Mol Biol* 2001, **312**:289-307.

54. •• Voigt CA, Mayo SL, Arnold FH, Wang ZG: **Computational method to reduce the search space for directed protein evolution.** *Proc Natl Acad Sci USA* 2001, **98**:3778-3783.
The authors used a mean field theory to determine each residue's structural tolerance to mutations. This tolerance is quantified by the residue's local sequence entropy, which is a measure of the effective number of amino acids that are structurally permitted at that site. For an *in vitro* directed evolution experiment, the authors suggest that mutations that enhance stability or activity are most likely to accumulate in these high entropy regions. Multiple compensating mutations are rare in such experiments, so mutations are most likely at sites that tolerate multiple amino acids. Calculations involving subtilisin E and T4 lysozyme are consistent with the mutations observed in directed evolution experiments.

55. Voigt CA, Mayo SL, Arnold FH, Wang ZG: **Computationally focusing the directed evolution of proteins.** *J Cell Biochem* 2001:58-63.

56. Sander C, Schneider R: **Database of homology-derived protein structures and the structural meaning of sequence alignment.** *Proteins* 1991, **9**:56-68.

57. •• Hill DJ, Mio MJ, Prince RB, Hughes TS, Moore JS: **A field guide to foldamers.** *Chem Rev* 2001, **101**:3893-4011.
A comprehensive review of nonbiological folding molecules.